IRIS DATASET EXPLORATION

CREATED BY RAJARSHI GUHA

PROBLEM STATEMENT

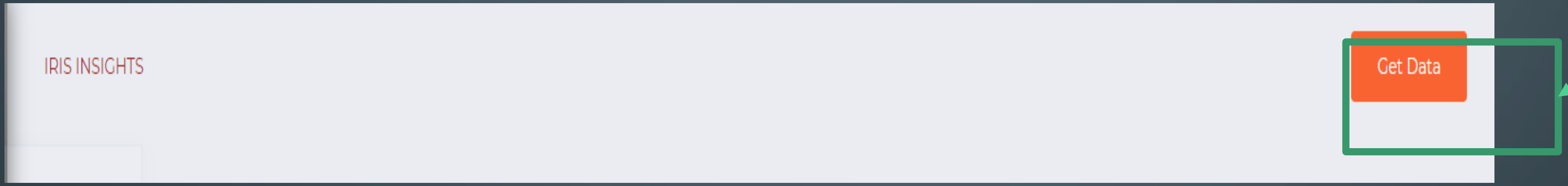
- A garden owner would like to examine the distinctiveness of different Iris flower classes (Iris Setosa, Iris Versicolour, and Iris Virginica) in his garden based on historical data measurements (sepal length, sepal width, petal length, petal width). With understanding of the differences, he'd also like to create a tool with help from a top data scientist to quickly retrieve records of the most similar Iris flowers in his garden for any input Iris flower.

SOLUTION

- Step 1:
  - Fetching the data . Most well known python machine learning libraries provide the iris dataset out of box. However here we will be fetching the data from the original source :https://archive.ics.uci.edu/ml/datasets/Iris
  - The application using Flask in python. All the visualizations are done in Bokeh.
  - There are 4 parameters:
    - Sepal width
    - Sepal length
    - Petal width
    - Petal length
  - The 3 species are : Iris-Setosa , iris-versicolor ,Iris-virginica
  - Each class is equally balanced with 50 observations is each

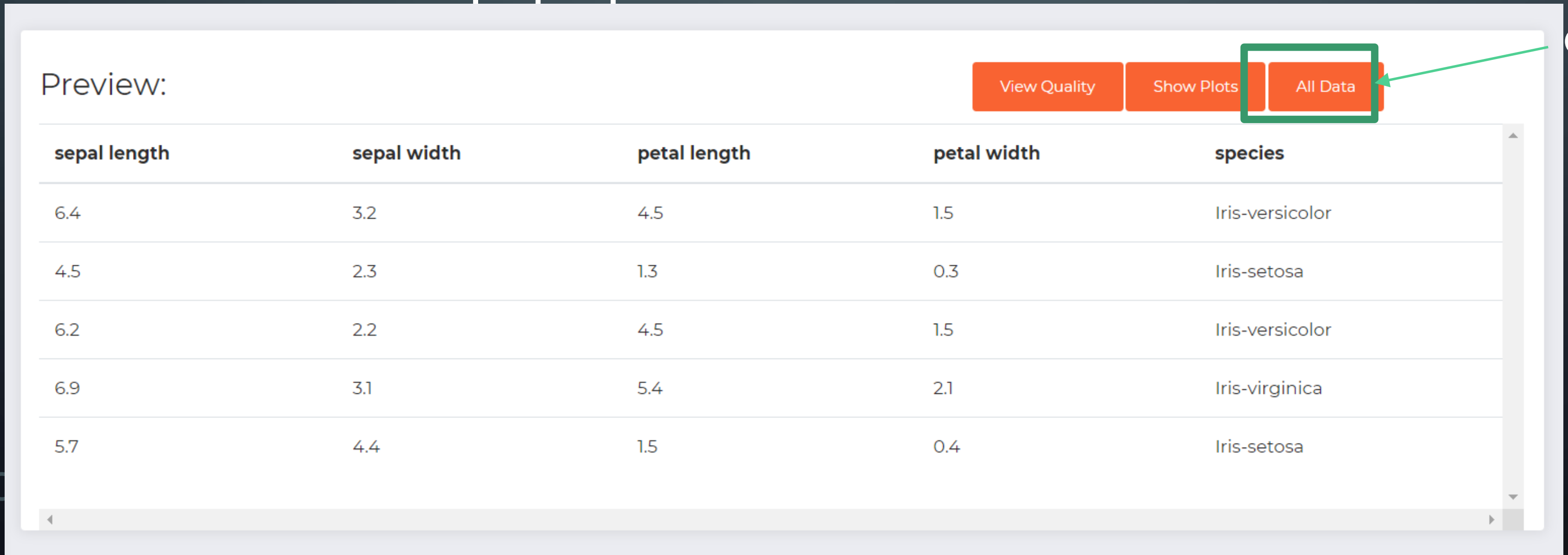
FETCHING DATA

- Install the app as mentioned in “Steps to run iris exploration app.docx” .
- Navigate to http://localhost:5000/dashboard



Click Here

- The result should pop up as shown below:



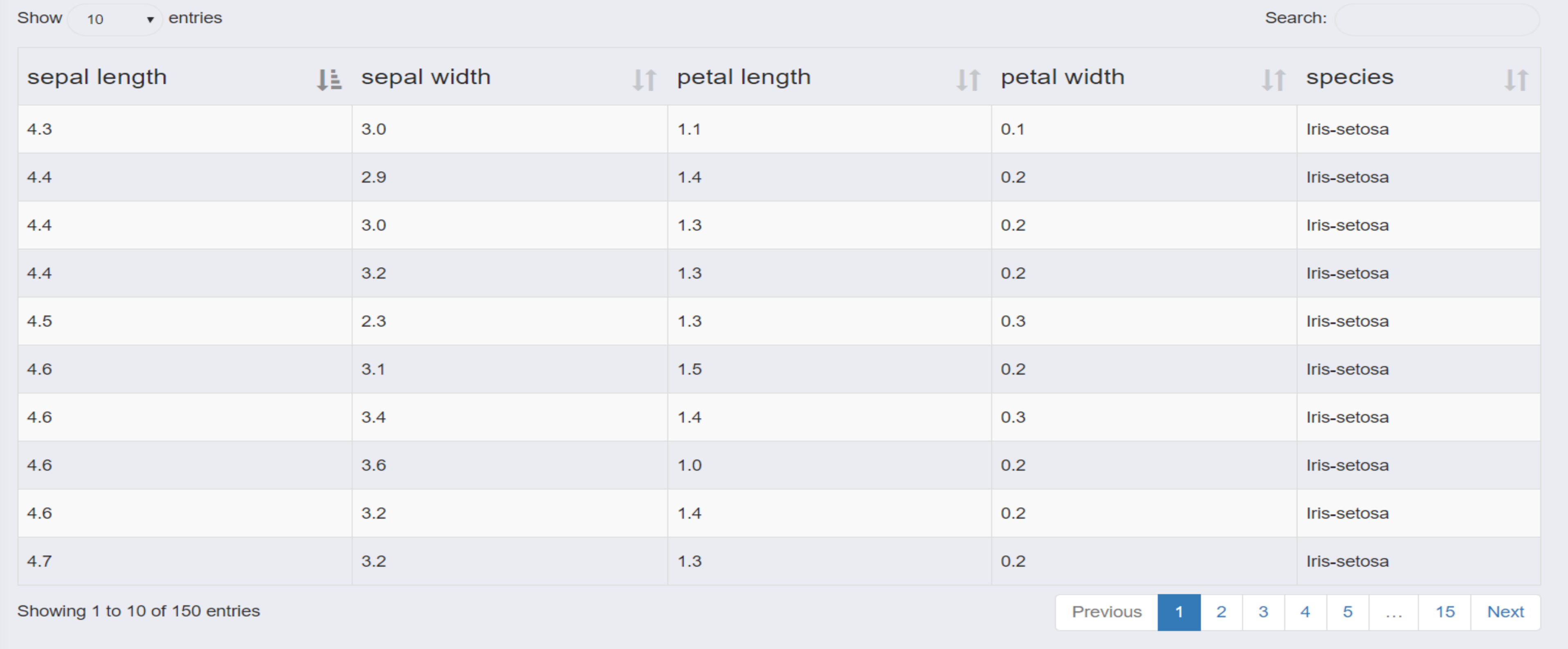
Preview:

sepal length	sepal width	petal length	petal width	species
6.4	3.2	4.5	1.5	Iris-versicolor
4.5	2.3	1.3	0.3	Iris-setosa
6.2	2.2	4.5	1.5	Iris-versicolor
6.9	3.1	5.4	2.1	Iris-virginica
5.7	4.4	1.5	0.4	Iris-setosa

Click Here

FETCHING DATA

- The user has the option to view the full dataset on clicking the “All data” button” as shown below:

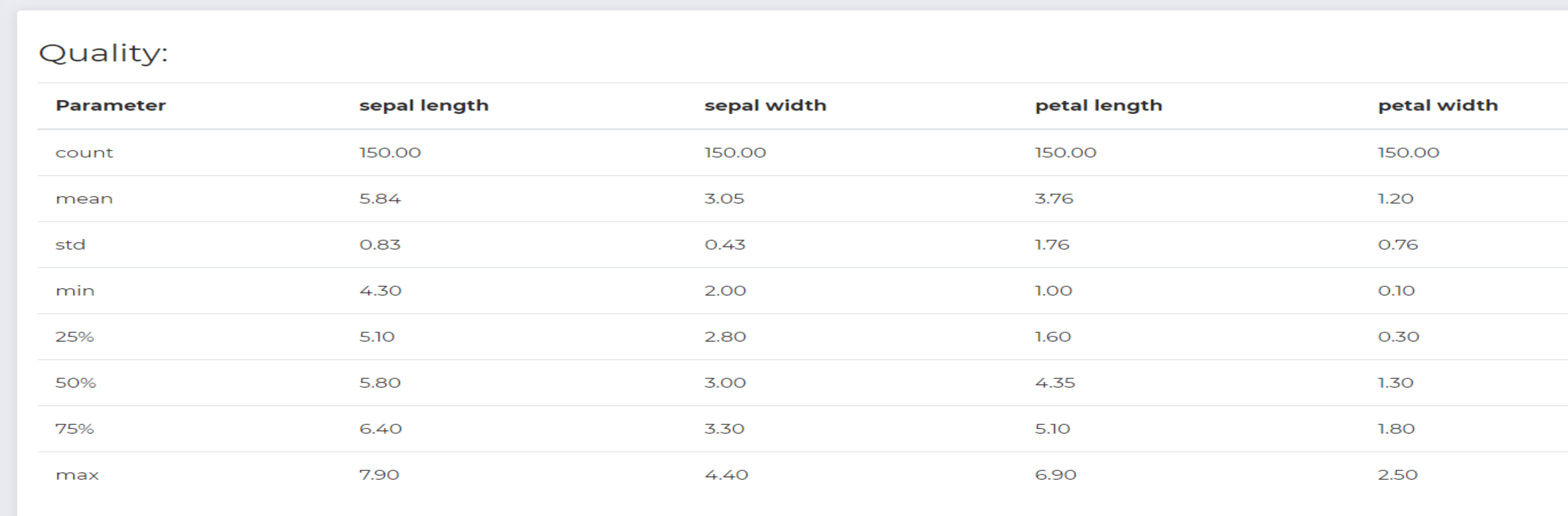


sepal length	sepal width	petal length	petal width	species
4.3	3.0	1.1	0.1	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.4	3.0	1.3	0.2	Iris-setosa
4.4	3.2	1.3	0.2	Iris-setosa
4.5	2.3	1.3	0.3	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
4.6	3.6	1.0	0.2	Iris-setosa
4.6	3.2	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa

Showing 1 to 10 of 150 entries

DATA QUALITY

- The backend send a request to the dataset URL to fetch the data which is then parsed and stored in a pandas data frame for easy slicing and dicing for our purposes.
- On clicking the "View Quality" button we get a result as shown below:



Quality:

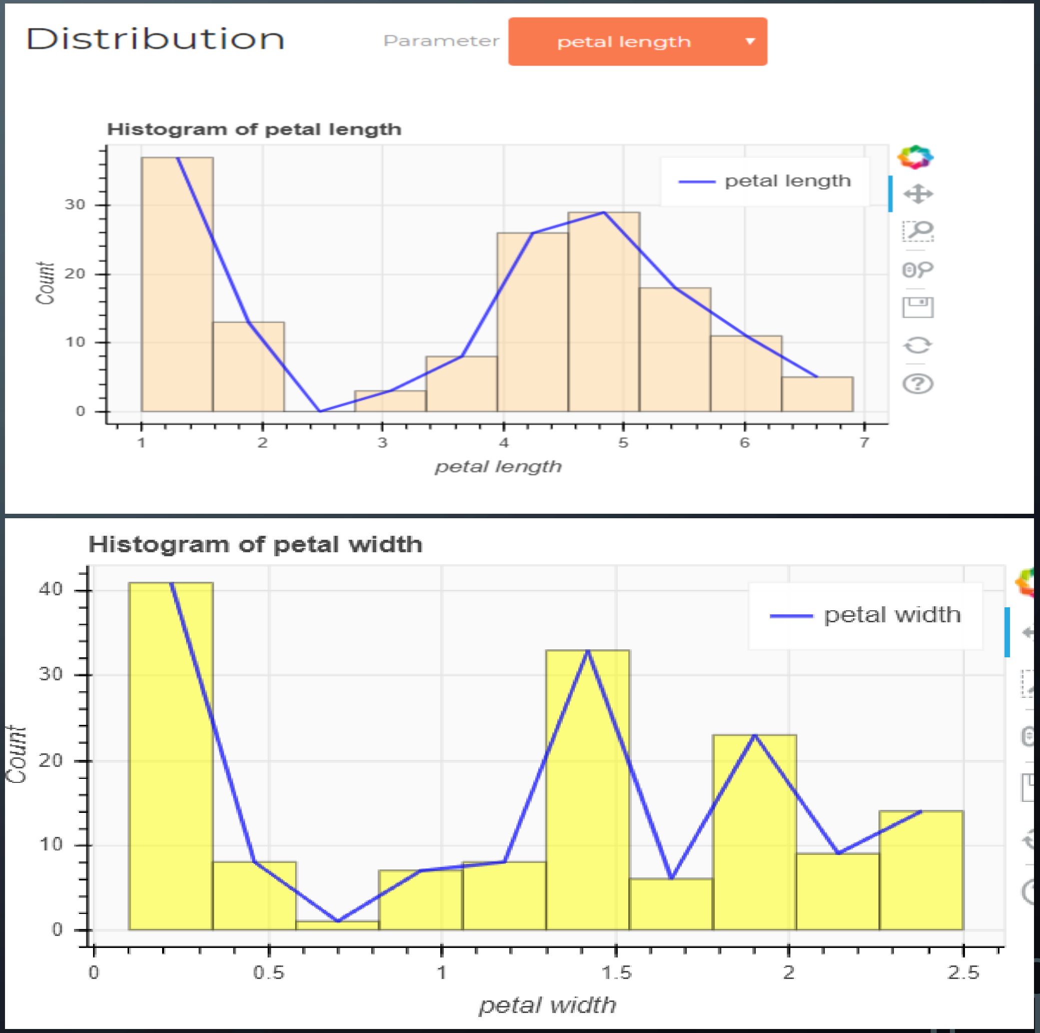
Parameter	sepal length	sepal width	petal length	petal width
count	150.00	150.00	150.00	150.00
mean	5.84	3.05	3.76	1.20
std	0.83	0.43	1.76	0.76
min	4.30	2.00	1.00	0.10
25%	5.10	2.80	1.60	0.30
50%	5.80	3.00	4.35	1.30
75%	6.40	3.30	5.10	1.80
max	7.90	4.40	6.90	2.50

DATA QUALITY

- The data quality gives us a very concise yet informative view of the dataset.
  - The count for all the observations equals 150. So we can rule out the presence of missing values.
  - The sepal length is definitely larger than the other parameters with a std deviation of 0.83 (~14%)
  - Sepal width has a similar std deviation as a % of the mean
  - Petal length and width have high standard deviation at ~46%. These features should then account for a large variation in the dataset which we shall investigate further. Comparing the standard deviation with the median provides similar results.
  - There seem to be some outliers in the dataset.

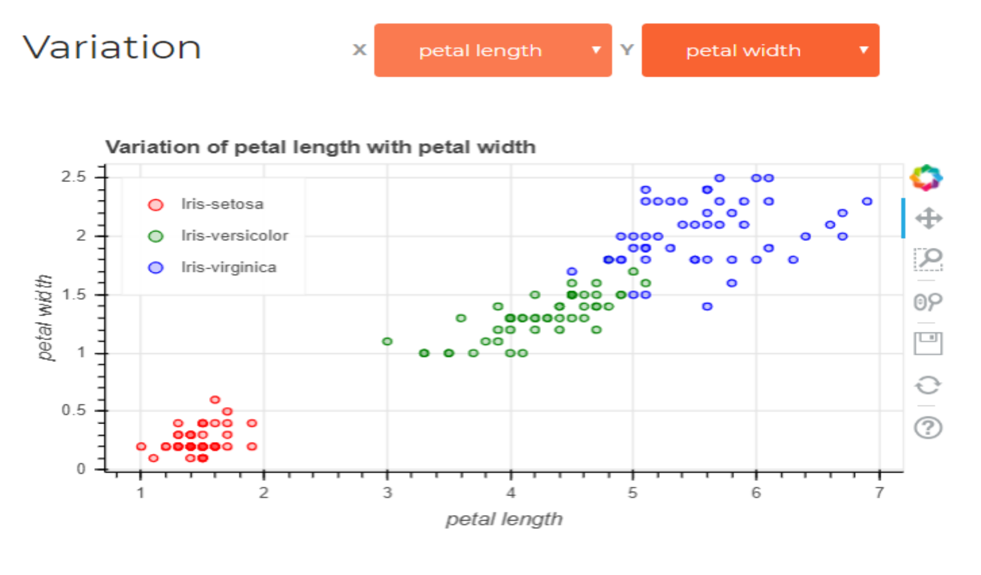
DATA EXPLORATION

- Click on the “Show Plots” button to show the plots available for the user in this dashboard.
  - Most of the petal lengths are between 1 &1.15 ~26% and between 4&6 ~36%
  - Most of the petal widths are between 0 &0.25 ~26% and between 1.5&2 ~36%



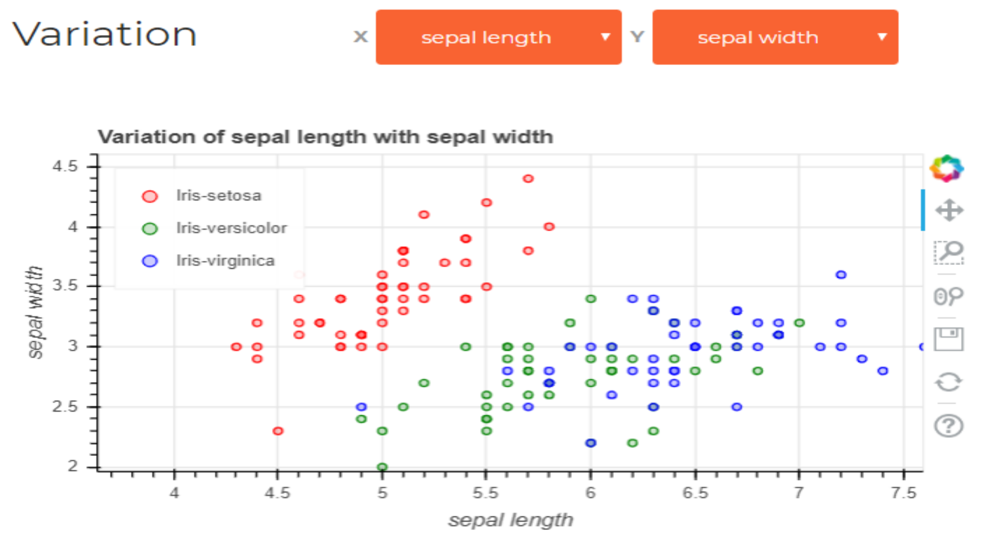
DATA EXPLORATION

- There seems to be a strong correlation between petal width and petal length given the slope of the best fit line.
- The 3 species of flowers are well separated in their petal width vs petal length distribution. This should aid in the classification of any new species of flower based on its parameters.
- In the case of sepal width vs sepal length the “Iris-setosa” is distinctly distributed to draw a classification but the same cannot be said about “Iris-versicolor” and Iris-verginica.



Variation

X petal length ▾ Y petal width ▾



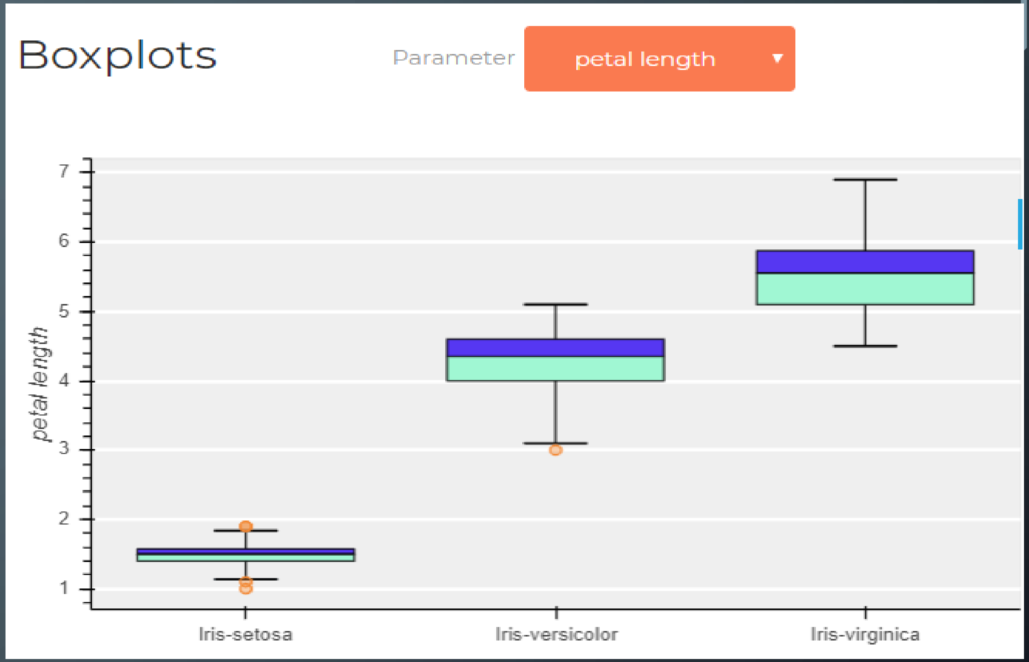
Variation

X sepal length ▾ Y sepal width ▾

DATA EXPLORATION

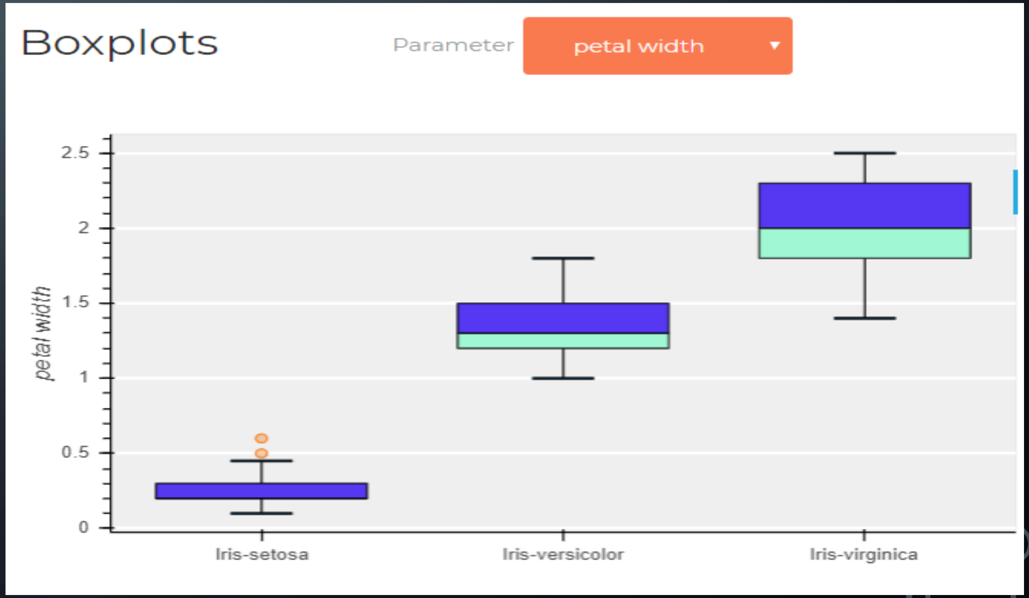
- Outliers present for both petal width and petal length for the iris-setosa species of flowers.

- Very small variance in the data for the Iris-setosa species of flowers in both petal length and petal width



Boxplots

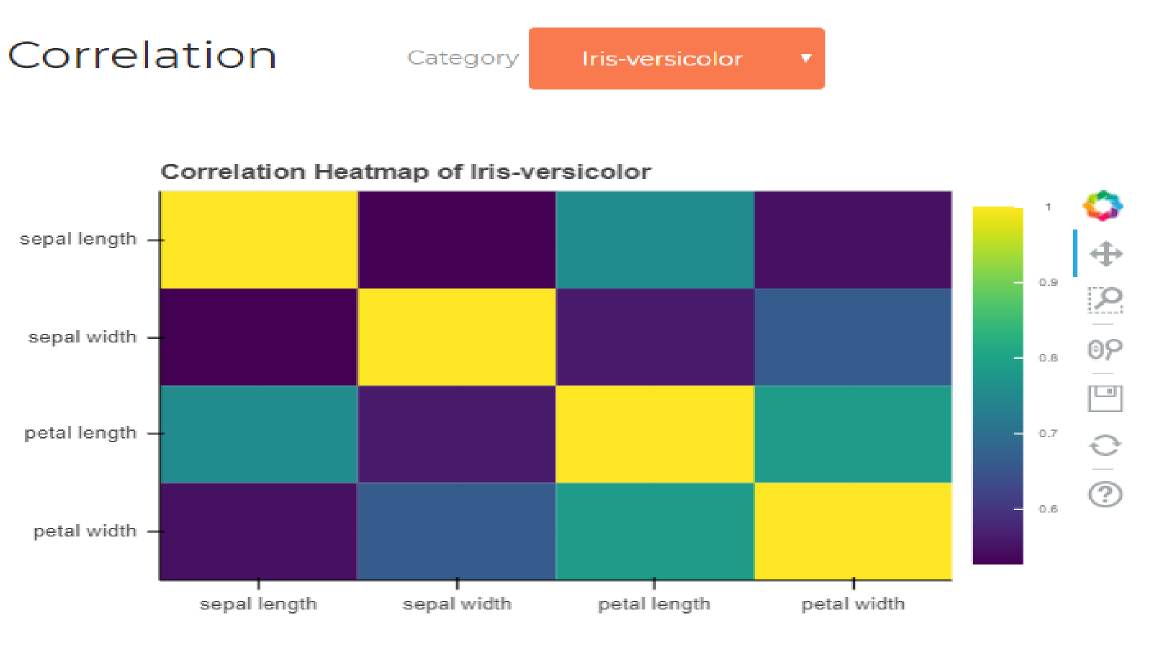
Parameter petal length



Boxplots

Parameter petal width

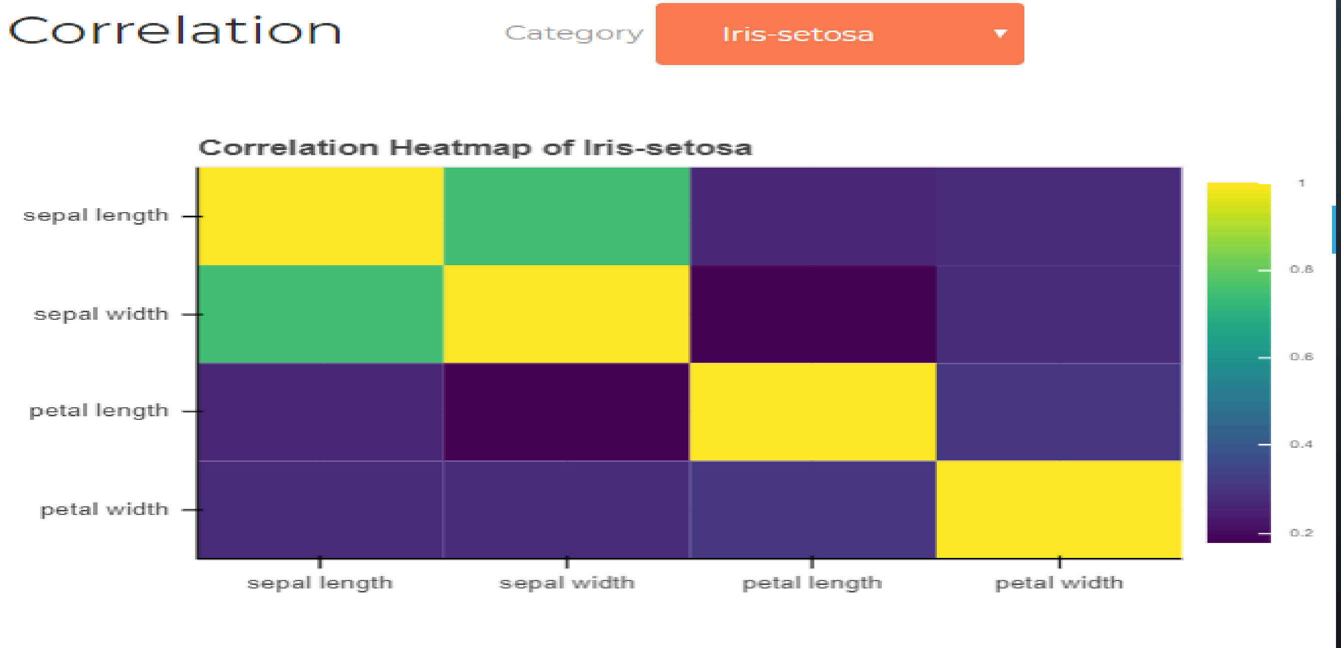
DATA EXPLORATION



Correlation

Category: Iris-versicolor

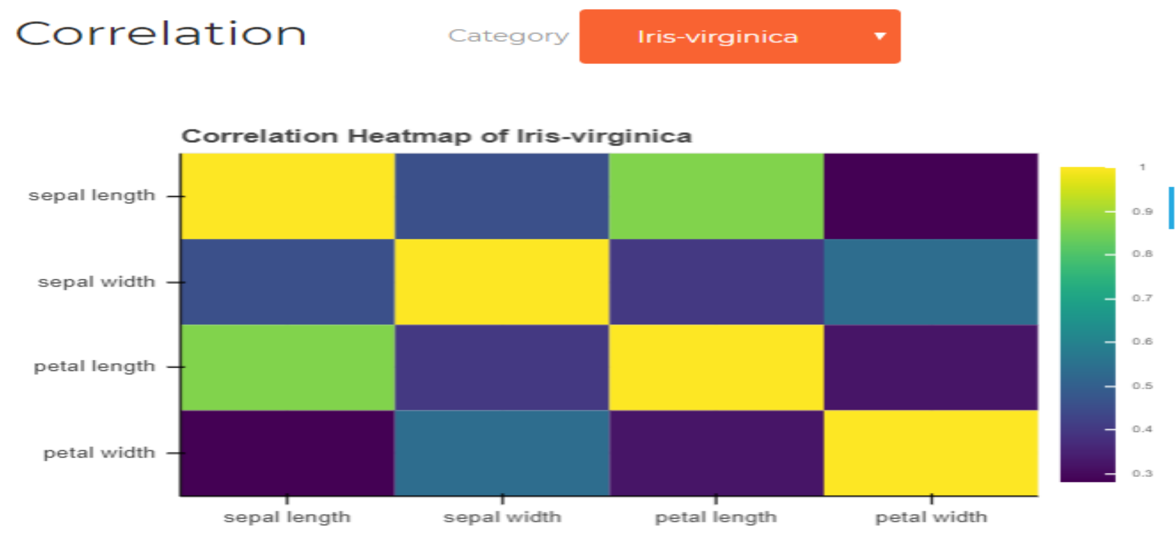
Correlation Heatmap of Iris-versicolor



Correlation

Category: Iris-setosa

Correlation Heatmap of Iris-setosa



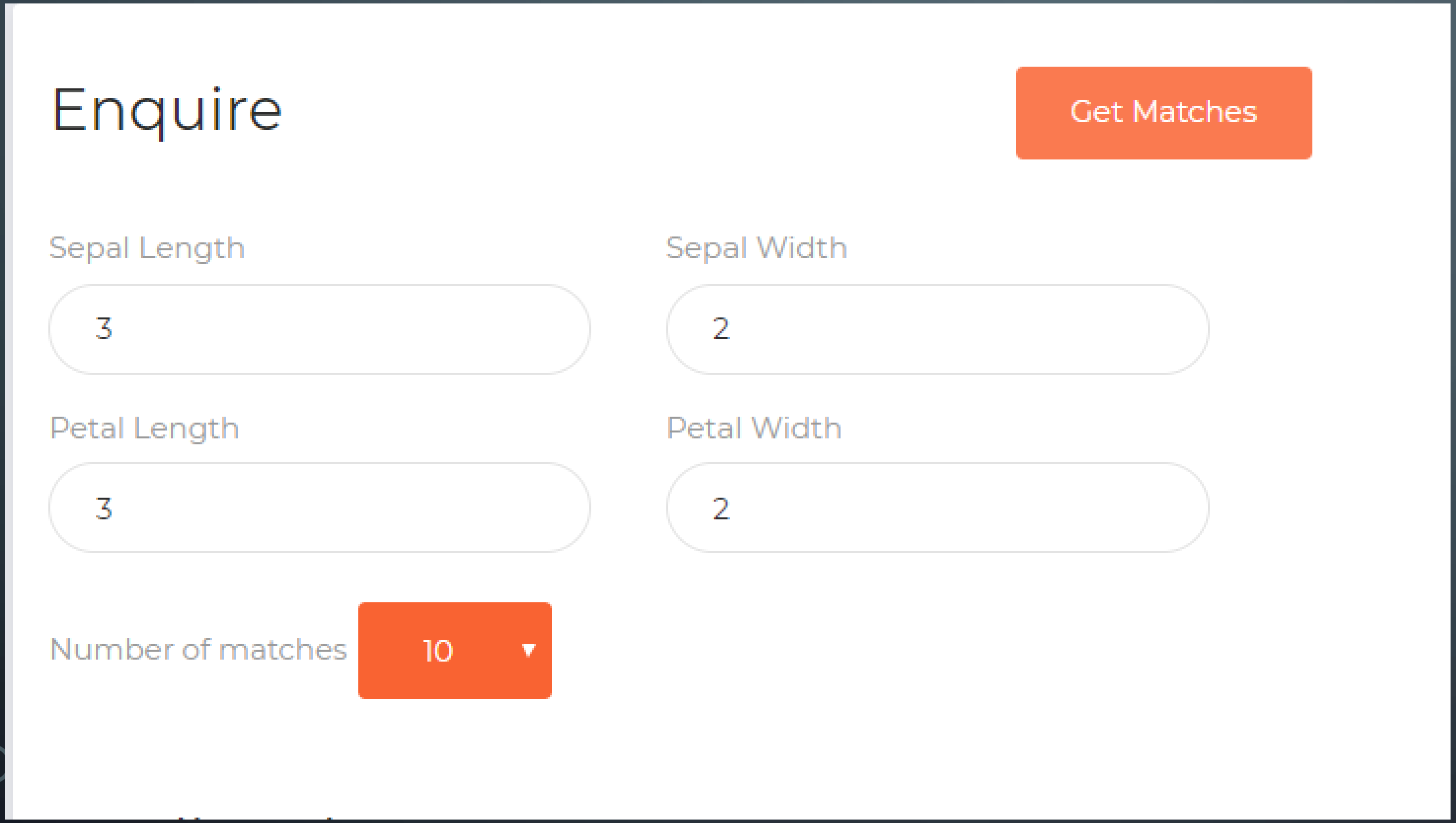
Correlation

Category: Iris-virginica

Correlation Heatmap of Iris-virginica

- Iris-Versicolor: Strong Correlation petal width and petal length
- Iris-Setosa: Strong Correlation sepal width and sepal length
- Iris-Virginica: Strong Correlation petal length and sepal length

ENQUIRY



Enquire

Sepal Length: 3

Sepal Width: 2

Petal Length: 3

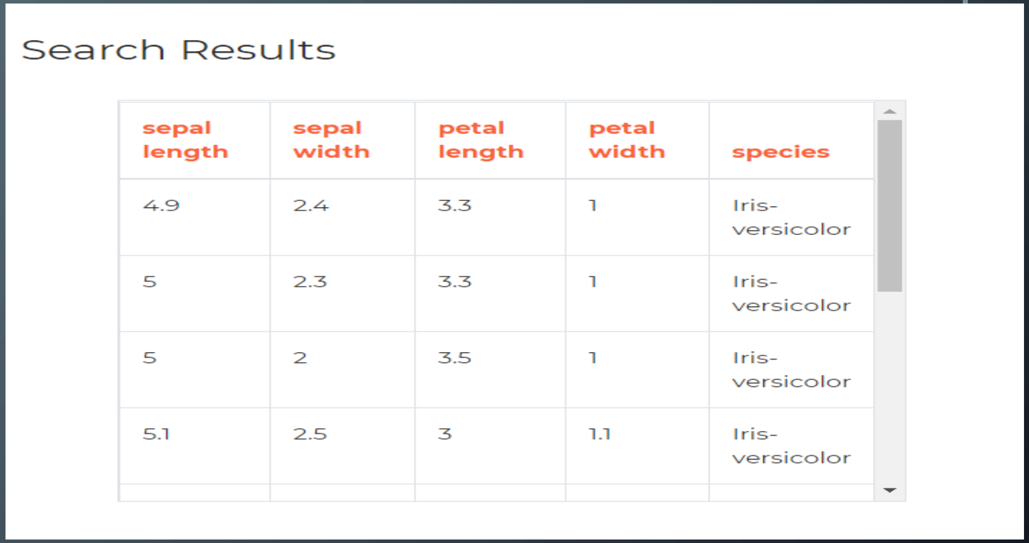
Petal Width: 2

Number of matches: 10

- Enter the inputs to search the closest matches.
- The app calculates an Euclidean distance between the search inputs and the observations in the iris data to return the closest matches.

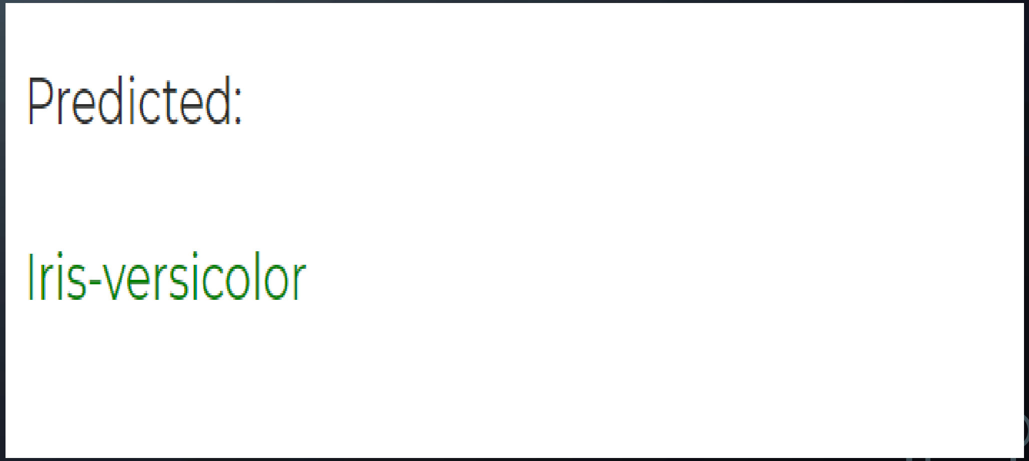
RESULTS

- The results are presented in the form of a table arranged in descending order of matching. The closest matches are ranked first .
- The application also predicts the species of the flower based on the inputs.
- The model used to predict the flower species is a Support Vector Classifier
- We obtain the first 2 principal components of the data to train the model with an accuracy of ~93%
- The model is prebuilt and served on the fly from the flask App itself.



Search Results

sepal length	sepal width	petal length	petal width	species
4.9	2.4	3.3	1	Iris-versicolor
5	2.3	3.3	1	Iris-versicolor
5	2	3.5	1	Iris-versicolor
5.1	2.5	3	1.1	Iris-versicolor



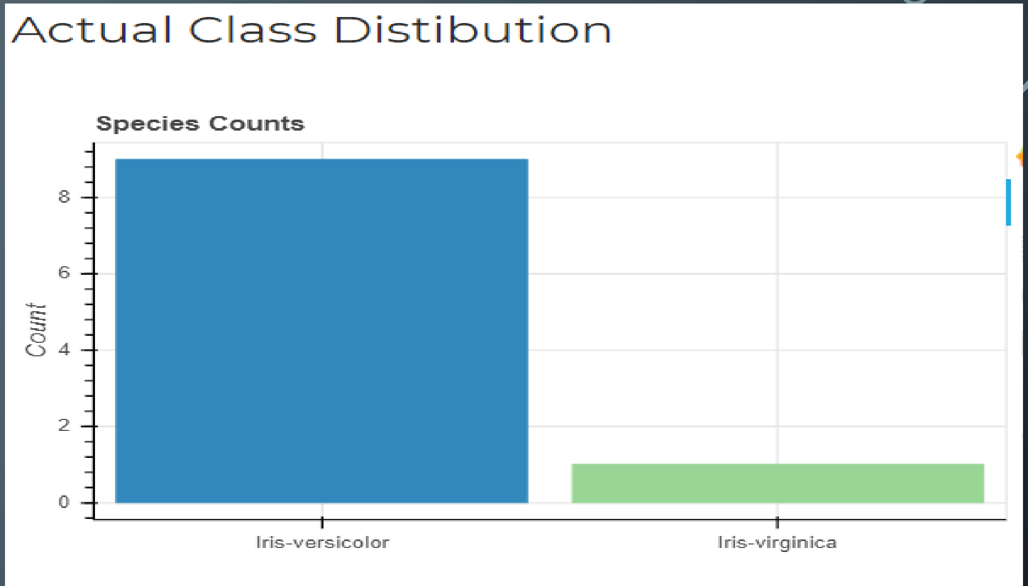
Predicted:

Iris-versicolor

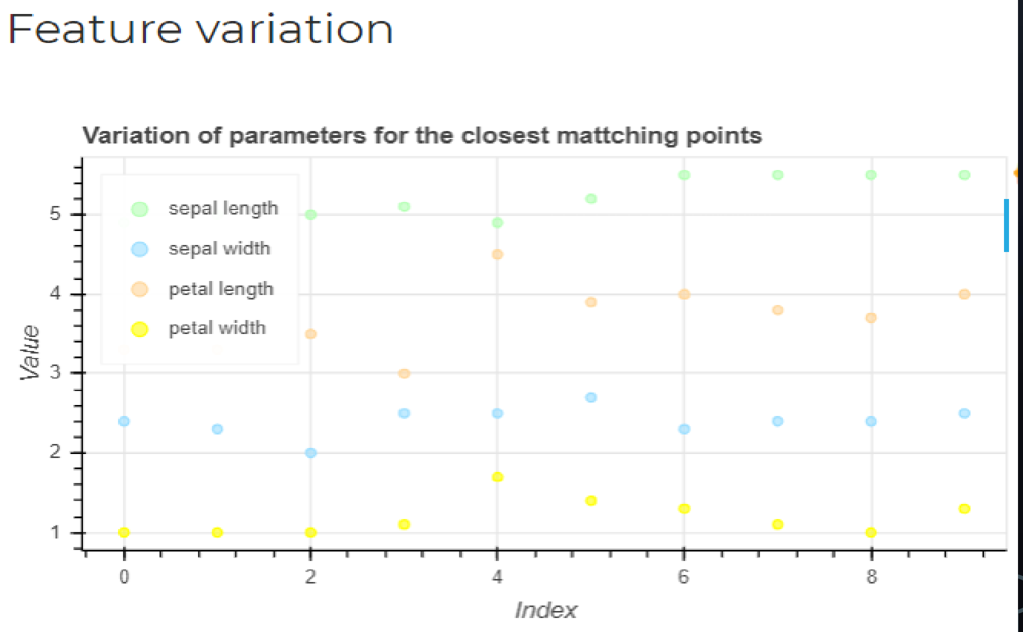
RESULTS

- The closest results retrieved from iris data have a class label.
- The Actual class distribution vertical stack plot helps the user to compare the the predicted class with the class present in the data

- Feature variation plot shows the variation of each feature in the retrieved closest matches.



Actual Class Distibution



Feature variation

IMPROVEMENTS

- Ability to make all the content in the page interactive including the plots
- Adding a login module and setting up user profiles to customize dashboards according to the user's role.
- Adding the ability to train and deploy models from the dashboard itself.